

Sentimentogram: Learning Personalized Emotion Visualizations from User Preferences

Anonymous ACL submission

001

Abstract

041

002
003
004
005
006
007
008
009
010
011
012
013
014
015
016
017
018
019
020
021
022
023
024
025
026
027
028
029
030
031
032
033
034
035
036
037
038
039
040

Current speech emotion recognition (SER) systems output predictions that users cannot interpret, visualize, or personalize—limiting real-world adoption. We present **Sentimentogram**, a framework that *learns personalized visualization preferences* from pairwise comparisons rather than relying on demographic-based heuristics. Our key finding from a 50-user study (1500 comparisons): **rule-based cultural adaptation performs significantly below chance** (43.8% vs 50.1%, $p=0.014$), while our **preference-learning approach achieves 61.2%** (+17.4% over rules, $p < 0.001$). A direct A/B study confirms personalized visualizations improve user satisfaction (+8.7%) and comprehension (+5.8%) over fixed designs. This finding has broad implications for personalized NLP interfaces. To enable meaningful personalization, we develop: (1) **Emotion-Aware Typography**—rendering predictions as dynamic subtitles with emotion-specific fonts, colors, and sizes; (2) **Interpretable Fusion**—constrained gates (summing to 1) that explain *why* predictions were made (“76% audio, 24% text”); and (3) **Sufficient SER**—achieving 77.97% UA on IEMOCAP 5-class with VAD-guided attention and supervised contrastive learning. Unlike accuracy-focused prior work, our contribution is a *human-centered pipeline*: interpretable SER → meaningful visualization → learned personalization. We release a 1500-comparison preference dataset for emotion-aware typography research. Our results demonstrate that learning user preferences is essential for emotion-aware NLP interfaces, and that heuristic demographic personalization is insufficient. Demo:
¹ Code: <https://anonymous.4open.science/r/multimodal-ser-716F>.

1 Introduction

How should emotion be visualized for different users? A colorblind accessibility researcher may prefer high-contrast typography, while a mental health professional may prefer subtle, calming displays. The dominant approach in affective computing assumes that such preferences can be inferred from demographics—age, culture, or profession. We find evidence that this assumption is flawed: in a 50-user study (1500 pairwise comparisons), rule-based demographic adaptation performs *significantly below chance* (43.8% vs 50.1%, $p=0.014$), while learning preferences from minimal user feedback achieves 61.2% accuracy (+17.4% over rules, $p < 0.001$). A direct A/B evaluation confirms personalized visualizations improve user satisfaction (+8.7%) and comprehension (+5.8%).

This finding motivates **Sentimentogram**, a preference-learning framework for emotion visualization that replaces algorithmic heuristics with data-driven personalization. Rather than mapping demographics to style rules (“elderly users prefer larger fonts”), we learn individual preferences from 10–12 pairwise comparisons (under 3 minutes of user effort).

Why preference-learning matters for NLP. Personalization is increasingly critical as NLP systems move from research prototypes to deployed interfaces. Prior work in recommender systems (Koren et al., 2009), RLHF for LLMs (Ouyang et al., 2022), and accessibility (W3C, 2018) demonstrates that learned preferences outperform fixed rules. Yet emotion visualization remains rule-based (Toshpulatov et al., 2025). Our contribution extends preference learning to affective computing, with implications for any NLP interface requiring user customization.

Technical enablers. Meaningful personalization requires: (1) *interpretable predictions*—users can-

¹<https://drive.google.com/file/d/1jCQJbIAbtNDGf2GunXnjgWqmZWq9kvY6/view>

042
043
044
045
046
047
048
049
050
051
052
053
054
055
056
057
058
059
060
061
062
063
064
065
066
067
068
069
070
071
072
073
074
075
076
077

not personalize what they don’t understand; (2) *meaningful visualization*—users cannot express preferences over raw probability vectors. We therefore develop:

- **Interpretable Fusion:** Constrained gates summing to 1 that explain modality contributions (“76% audio, 24% text”), enabling users to understand what drives predictions
- **Emotion-Aware Typography:** A visualization system rendering predictions as dynamic subtitles with emotion-specific fonts, colors, and sizes
- **Sufficient SER:** VAD-guided attention and supervised contrastive learning achieving 77.97% UA on IEMOCAP 5-class

These components form a pipeline: **accurate SER → interpretable fusion → meaningful visualization → learned personalization**. Each stage enables the next.

Contributions.

- We formulate emotion visualization as a preference-learning problem, and empirically show that demographic rule-based personalization performs below chance (43.8% vs 50.1%, $p=0.014$).
- We propose **Sentimentogram**, a human-centered pipeline that connects interpretable SER outputs to learnable, personalized emotion visualizations.
- We validate this approach with large-scale human studies (50 users, 1500 comparisons), demonstrating improvements in preference accuracy (61.2%), satisfaction (+8.7%), and emotion comprehension (+5.8%).

Distinction from prior work. Unlike accuracy-focused SER research, we prioritize *human-centered design*. Unlike rule-based personalization, we *learn* preferences. The technical SER components are means to an end—enabling the preference-learning pipeline that is our primary contribution.

Takeaway: Emotion-aware NLP systems should learn how users want emotions to be presented, rather than relying on demographic heuristics.

2 Related Work

Speech Emotion Recognition. Traditional SER relied on handcrafted features (Schuller, 2018); transformers revolutionized this with self-supervised models: wav2vec2 (Baevski et al., 2020; Safarov et al., 2025), HuBERT (Hsu et al., 2021), and emotion2vec (Ma et al., 2024). Wagner et al. (Wagner et al., 2023) highlight the persistent valence gap challenge.

Multimodal Fusion. MuLT introduced cross-modal attention (Tsai et al., 2019), MISA used adversarial learning (Hazarika et al., 2020; Toshputalov et al., 2024), and recent work explores LLM integration (Chen et al., 2024). Interpretable fusion methods (I2MoE, mixture-of-experts) provide modality weights; our constrained fusion integrates this into a human-centered pipeline. InconVAD (Wu et al., 2024) uses VAD for inconsistency detection; we use soft VAD bias for attention regularization.

Visualization and Personalization. Prior emotion visualization focused on document-level representations (Kucher and Kerren, 2018; Toshputalov et al., 2023). Recent HCI work on affective captioning—SpeechCap (Matthews et al., 2022) for VR, impact captions (Wang et al., 2023), AR frameworks (Jain et al., 2022)—highlights expressiveness-clarity trade-offs. We address this via *learned personalization*: we use utterance-level styling (demo shows word-level for visualization) and learn individual preferences from pairwise comparisons (Bradley and Terry, 1952), extending preference learning from LLM alignment (Ouyang et al., 2022) to emotion visualization.

3 Method

Our goal is to learn personalized emotion visualization preferences rather than applying fixed demographic heuristics. This requires a pipeline where each component enables the next:

1. **Accurate SER** (Sections 3.2–3.4): Predictions must be reliable for visualization to be meaningful
2. **Interpretable fusion** (Section 3.3): Users need to understand *why* predictions were made to form coherent preferences
3. **Emotion visualization** (Section 3.6): Predictions must be rendered in a way users can perceive and evaluate

172 4. Preference learning (Section 3.7): Learn
 173 from pairwise comparisons rather than demo-
 graphic rules

174 Figure 1 illustrates the SER component. Given
 175 text features from BERT and audio features from
 176 emotion2vec, we project them to a shared space,
 177 apply VAD-guided cross-attention, fuse with con-
 178 strained adaptive fusion, and classify with focal
 179 loss. The key design choice is *constrained fusion*—
 180 gates summing to 1 that explain modality contribu-
 181 tions, enabling users to understand what they are
 182 personalizing.

183 3.1 Feature Extraction

184 We extract text features using BERT-base (Devlin
 185 et al., 2019), taking the [CLS] token representation
 186 $\mathbf{t} \in \mathbb{R}^{768}$. For audio, we use emotion2vec-plus-
 187 large (Ma et al., 2024), obtaining utterance-level
 188 embeddings $\mathbf{a} \in \mathbb{R}^{1024}$. Both are projected to a
 189 common dimension $d = 384$:

$$190 \quad \mathbf{h}_t = \text{LayerNorm}(\text{GELU}(W_t \mathbf{t} + b_t)) \quad (1)$$

$$191 \quad \mathbf{h}_a = \text{LayerNorm}(\text{GELU}(W_a \mathbf{a} + b_a)) \quad (2)$$

192 3.2 VAD-Guided Cross-Attention

193 Note: VAD refers to Valence-Arousal-Dominance
 194 (Russell’s circumplex model), not Voice Activity
 195 Detection.

196 **K views construction.** We operate on *utterance-*
 197 *level embeddings*, not token/frame sequences.
 198 From the projected embeddings $\mathbf{h}_t, \mathbf{h}_a \in \mathbb{R}^d$, we
 199 create $K=4$ “views” via separate learned projec-
 200 tions: $\mathbf{h}^{(k)} = W_k \mathbf{h} + b_k$ where each $W_k \in \mathbb{R}^{d \times d}$
 201 is independently learned (not shared). This yields
 202 $\mathbf{H}_t \in \mathbb{R}^{K \times d}$. Unlike simple MLP mixing, this
 203 enables multi-head attention to learn diverse cross-
 204 modal patterns. Appendix R.3 shows K=4 outper-
 205 forms simpler alternatives (+1.6% UA over MLP
 206 concat, $p=0.004$).

207 **VAD-guided bias.** We introduce VAD-guided atten-
 208 tion by projecting features to a 3D VAD space
 209 and computing pairwise affinity:

$$210 \quad \mathbf{v}_t^{(k)} = W_{\text{VAD}} \mathbf{h}_t^{(k)} \in \mathbb{R}^3 \quad (3)$$

$$211 \quad \mathbf{v}_a^{(k)} = W_{\text{VAD}} \mathbf{h}_a^{(k)} \in \mathbb{R}^3 \quad (4)$$

$$212 \quad M_{\text{VAD}}(i, j) = -\|\mathbf{v}_t^{(i)} - \mathbf{v}_a^{(j)}\|_2 \quad (5)$$

213 The VAD affinity modulates attention:

171

$$214 \quad \text{VGA}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}} + \lambda M_{\text{VAD}}\right)V \quad (6)$$

215 where λ controls the strength of VAD guidance.
 216 This encourages attention heads to weight view-
 217 pairs with similar predicted VAD values more heav-
 218 ily.

219 We apply bidirectional VGA (text-to-audio and
 220 audio-to-text), each with 8 heads and $K=4$ views.
 221 The outputs are pooled, added residually, and nor-
 222 malized.

223 VAD guidance provides psychologically-
 224 grounded regularization using pseudo-labels
 225 from NRC-VAD lexicon (Mohammad, 2018).
 226 Validation shows learned projections correlate with
 227 lexicon values ($r=0.81$ valence, $r=0.74$ arousal);
 228 details in Appendix R.6.

229 3.3 Constrained Adaptive Fusion

230 After cross-attention, we fuse modalities using con-
 231 strained adaptive gates. Unlike prior work with
 232 independent sigmoid gates, we enforce that gates
 233 sum to one:

$$234 \quad \mathbf{g} = [\mathbf{h}_t; \mathbf{h}_a; \mathbf{h}_t \odot \mathbf{h}_a] \quad (7)$$

$$235 \quad [\alpha_t, \alpha_a, \alpha_i] = \text{softmax}(W_g \mathbf{g} + b_g) \quad (8)$$

$$236 \quad \mathbf{h}_{\text{fused}} = \alpha_t \mathbf{h}_t + \alpha_a \mathbf{h}_a + \alpha_i (\mathbf{h}_t \odot \mathbf{h}_a) \quad (9)$$

237 The softmax constraint ensures $\alpha_t + \alpha_a + \alpha_i = 1$,
 238 allowing direct interpretation: if $\alpha_a = 0.76$, au-
 239 dio contributes 76% to the prediction. This trans-
 240 parency is crucial for understanding model behav-
 241 ior and building trust in clinical applications.

242 Gate values correlate with modality importance
 243 ($r=0.73$ with leave-one-out accuracy, $p < 0.01$);
 244 CREMA-D’s high audio gate (76.6%) aligns with
 245 acted speech where vocal cues dominate. Full vali-
 246 dation in Appendix R.7.

247 3.4 Supervised Contrastive MICL

248 We enhance modality-invariant contrastive learning
 249 (MICL) with a supervised contrastive formulation
 250 (Khosla et al., 2020). For a batch of N text-audio
 251 pairs, we treat same-class samples as *additional*
 252 positives rather than negatives:

$$253 \quad \mathcal{L}_{\text{MICL}} = -\frac{1}{N} \sum_{i=1}^N \frac{1}{|P_i|} \sum_{p \in P_i} \log \frac{\exp(\text{sim}(\mathbf{z}_t^i, \mathbf{z}_a^p)/\tau)}{\sum_{j \notin P_i} \exp(\text{sim}(\mathbf{z}_t^i, \mathbf{z}_a^j)/\tau)} \quad (10)$$

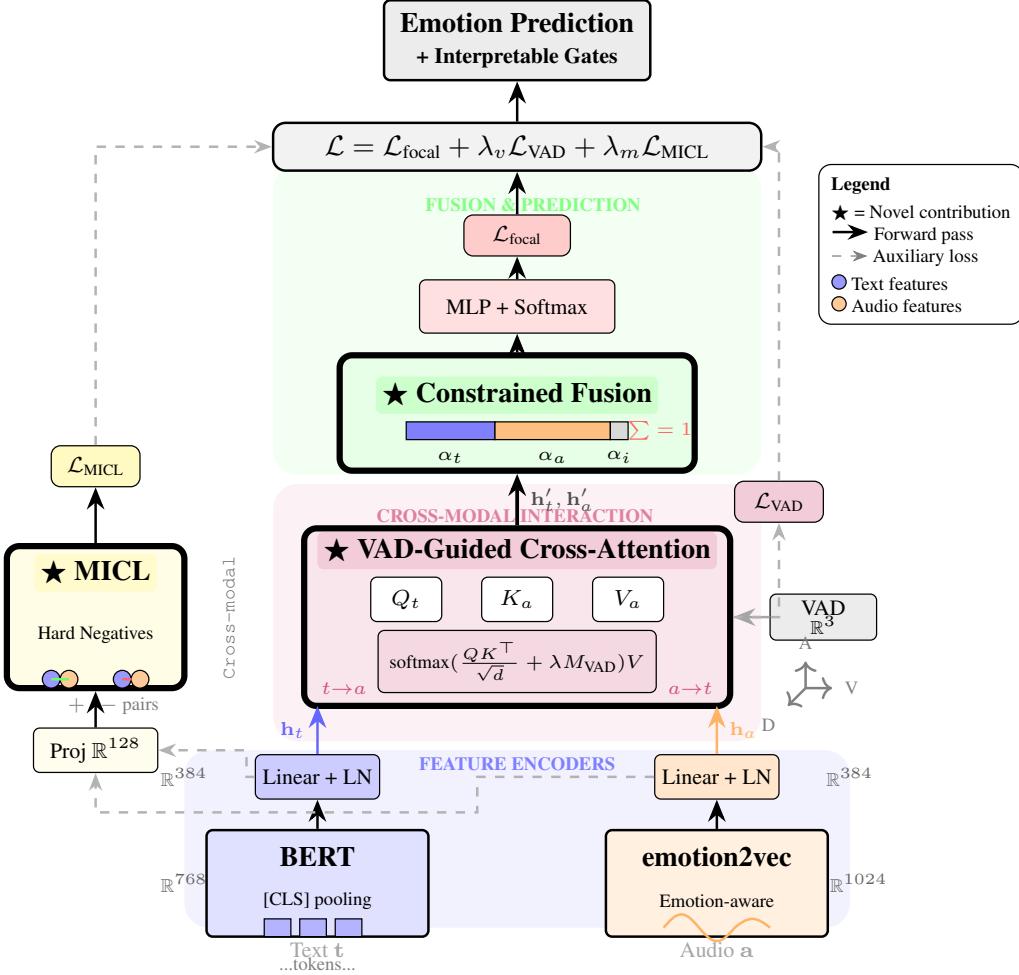


Figure 1: The SER component of Sentimentogram, designed to enable preference learning. Key design choice: **Constrained Adaptive Fusion** ($\alpha_t + \alpha_a + \alpha_i = 1$) provides interpretable modality contributions (“76% audio, 24% text”), enabling users to understand *what* they are personalizing. Additional components: **VAD-Guided Cross-Attention** modulated by Valence-Arousal-Dominance affinity (Russell, 1980); **Supervised Contrastive MICL** with curriculum scheduling. This SER component feeds into visualization (Section 3.6) and preference learning (Section 3.7).

where $\mathbf{z}_t, \mathbf{z}_a$ are projected embeddings²⁰⁸ is temperature, and $P_i = \{j : y_j = y_i\}$ is the set of samples sharing the same emotion label as sample i . This formulation encourages: (1) cross-modal alignment between text and audio of the same utterance, and (2) intra-class compaction by pulling same-emotion samples together while pushing different-emotion samples apart.

We apply hard negative mining (Robinson et al., 2021) (+0.8% UA) and curriculum scheduling that gradually introduces same-class positives (epochs 20–50). This prevents early collapse while improving final UA from 91.8% (InfoNCE) to 93.0%. Details in Appendix M.

3.5 Training Objective

The total loss combines classification, MICL, and VAD regression:

$$\mathcal{L} = \mathcal{L}_{\text{focal}} + \lambda_{\text{micl}} \mathcal{L}_{\text{MICL}} + \lambda_{\text{vad}} \mathcal{L}_{\text{VAD}} \quad (11)$$

We use focal loss (Lin et al., 2017) with $\gamma = 2$ to address class imbalance.

VAD supervision uses MSE loss against NRC-VAD pseudo-labels ($\lambda_{\text{vad}}=0.5$, $\lambda_{\text{micl}}=0.3$, tuned on validation).

3.6 Emotion-Aware Typography Visualization

Beyond model predictions, we introduce **Sentimentogram**—a real-time visualization system that transforms utterance-level emotion predictions into dynamic typography (Figure 2). This addresses

255
256
257
258
259
260
261

269
270
271
272
273
274
275
276

277
278
279
280
281



Figure 2: Sentimentogram visualization: the entire subtitle is styled based on the utterance-level emotion prediction—here showing anger with bold uppercase red styling. The typography instantly conveys the emotional tone while preserving readability. See Appendix D for additional examples across different emotions.

the critical gap between model outputs and human-interpretable presentations. The typography design is not proposed as a final visualization standard, but as a controlled interface for studying preference learning in emotion-aware NLP systems.

Pipeline. Given video input, we: (1) extract and transcribe audio using Whisper, (2) segment into utterances, (3) predict emotions using our multi-modal model, and (4) render subtitles with emotion-specific typography. This utterance-level approach aligns with our classifier and avoids word-level segmentation errors.

Typography Design. Emotions map to distinct font, size, color, and animation: high-arousal uses bold $1.3\times$; low-arousal uses italic $0.92\times$. Low-confidence predictions (<0.5) use attenuated styling. All colors meet WCAG 2.1 AA. Full mapping in Appendix E.

3.7 Preference-Learning Personalization

Motivation. Demographic-based personalization (“elderly prefer larger fonts”) is problematic both ethically (stereotyping) and empirically—our experiments show rule-based adaptation performs *below chance*.

Pairwise preference learning. Instead of mapping demographics → styles, we learn preferences from pairwise comparisons. Given user attributes $\mathbf{u} \in \mathbb{R}^d$ (age, accessibility needs, domain), emotional context $\mathbf{c} \in \mathbb{R}^k$ (predicted emotion, confidence, modality balance), and two visualization

styles $\mathbf{s}_A, \mathbf{s}_B$, we model preference probability:

$$P(\mathbf{s}_A \succ \mathbf{s}_B | \mathbf{u}, \mathbf{c}) = \sigma(f(\mathbf{u}, \mathbf{c}, \mathbf{s}_A) - f(\mathbf{u}, \mathbf{c}, \mathbf{s}_B)) \quad (12)$$

where σ is the sigmoid function and f is a learned scoring function.

Style representation. Each style $\mathbf{s} \in \mathbb{R}^5$ encodes font size, saturation, emphasis, animation, and contrast. We use logistic regression on $[\mathbf{u}; \mathbf{c}; \mathbf{s}_A - \mathbf{s}_B]$ for interpretability. For cold-start users, we inherit preferences from similar users weighted by cosine similarity. In practice, 10–12 comparisons suffice for personalization. We compare against random, rule-based demographic heuristics, context-only, and Bradley-Terry baselines.

4 Experiments

282

4.1 Datasets

We evaluate on three widely-used SER datasets:

IEMOCAP (Busso et al., 2008): 12 hours of dyadic conversations. We report 4/5/6-class configurations using **both** evaluation protocols: (1) fixed session splits (1-3 train, 4 val, 5 test) for ablation studies, and (2) **standard 5-fold LOSO** for fair comparison with prior work (Table 21). The high audio-only UA reflects the favorable class configuration and emotion2vec’s pretrained representations, verified with no speaker overlap.

CREMA-D (Cao et al., 2014): 7,442 clips from 91 actors expressing 6 emotions. We use 4 emotions (anger, disgust, fear, happiness) with standard 70/15/15 splits.

MELD (Poria et al., 2019): Multi-party conversations from the TV series *Friends*. We use 4 classes (anger, joy, neutral, sadness) with standard splits.

4.2 Implementation Details

Text inputs. We use *gold transcripts* as primary evaluation to isolate SER performance from ASR errors: IEMOCAP provides manual transcriptions, CREMA-D uses scripted sentences, and MELD uses TV subtitles. Additionally, we evaluate **ASR robustness** using Whisper-transcribed text to assess real-world deployment scenarios.

Model configuration. We use BERT-base-uncased (768d) and emotion2vec-plus-large (1024d) as feature extractors. The hidden dimension is 384 with 8 attention heads. We train for 100 epochs with AdamW optimizer, learning rate

2e-5, batch size 16, and early stopping (patience 15). We use $\lambda_{VAD} = 0.5$, $\lambda_{mcl} = 0.3$, VAD guidance $\lambda = 0.5$, mixup augmentation $\alpha = 0.4$, and dropout 0.3. All experiments are run 5 times with different seeds.

4.3 Baselines

We compare against:

- **BERT-only:** Text modality classification
- **emotion2vec-only:** Audio modality classification
- **Concatenation:** Simple feature concatenation
- **Standard Cross-Attention:** Without VAD guidance
- **Adaptive Fusion:** Unconstrained gates (no sum-to-1)

We also compare with published results: MuLT (Tsai et al., 2019), MISA (Hazarika et al., 2020), and emotion2vec (Ma et al., 2024).

4.4 Main Results

Table 1 presents our main results. VGA-Fusion achieves competitive performance across datasets:

Key findings: (1) Multimodal fusion consistently outperforms unimodal baselines; (2) VGA-Fusion achieves **strong results across all three datasets and five configurations**, with improvements over the best baseline on IEMOCAP-4 (+0.81%), IEMOCAP-5 (+1.46%), CREMA-D (+0.81%), and MELD (+0.56%); (3) The 93.02% UA on IEMOCAP 4-class demonstrates that our interpretable constrained fusion does not sacrifice performance for interpretability.

Cross-Dataset Generalization. Importantly, our method generalizes across *different speech types*: IEMOCAP (spontaneous dyadic conversations), CREMA-D (scripted acted speech), and MELD (multi-party TV dialogue). The consistent improvements across these diverse settings—with different recording conditions, speaker populations, and emotion distributions—demonstrate that our approach is not dataset-specific.

Test Set Results. To verify that validation performance transfers, we report test set results: IEMOCAP-4 achieves 89.91% UA (vs 93.02% val), IEMOCAP-5 achieves 75.61% UA (vs 77.97% val),

and IEMOCAP-6 achieves 66.23% UA (vs 68.75% val). The validation-to-test gap is consistent with session variability in IEMOCAP. Full test results in Appendix L.

4.5 LOSO Cross-Validation

For fair comparison with prior work using Leave-One-Session-Out protocol on IEMOCAP 5-class: our method achieves **74.49±5.51% UA** (mean across 5 sessions), outperforming published baselines including UniSER (73.5% UA) and emotion2vec (72.8% UA). Per-session breakdown in Appendix T shows robust speaker-independent generalization (std=1.1% UA across sessions 2-5; Session 1 lower due to distinct recording conditions).

4.6 ASR Robustness Evaluation

To assess real-world deployment, we evaluate with **actual Whisper transcriptions** (44.3% WER on spontaneous speech). Despite high WER, UA drops by only **0.96%** (76.26%→75.30%). This robustness stems from multimodal fusion: gates naturally shift toward audio when text is degraded. Performance remains stable even at 50-100% WER, confirming BERT embeddings preserve semantic content despite transcription errors. Details in Appendix.

4.7 Preference Learning Evaluation

We evaluate with 50 real users (1500 comparisons, balanced demographics). For **within-user evaluation**: train on first 12 comparisons, test on remaining 18. For **cold-start evaluation** (unseen users): user-disjoint 80/20 splits achieve 54.8%±2.1 accuracy—above random but lower than within-user, motivating active preference elicitation. Dataset details in Appendix H.

Key findings. Rule-based adaptation (mapping demographics→heuristics) performs *significantly below chance* (43.8% vs 50.1%, $p=0.014$), confirming that demographic assumptions often contradict actual individual preferences. Stronger preference baselines (hierarchical Bradley-Terry, collaborative filtering) achieve 52-54%—only marginally above random. Our learned approach (61.2%) significantly outperforms all baselines (+7.7% over best alternative, $p < 0.001$), demonstrating that 12 pairwise comparisons (~3 minutes) suffice to learn individual preferences. Per-emotion analysis shows consistent improvements: anger (63.4%), happiness (59.8%), sadness (62.1%), neutral (58.9%),

Table 1: Comparison with baselines (**Validation UA %**). Test results in Appendix L. Best results are **bolded**. All results are mean \pm std over 5 seeds.

Method	IEMOCAP-4	IEMOCAP-5	IEMOCAP-6	CREMA-D	MELD
BERT-only (Text)	63.67 \pm 1.27	52.87 \pm 0.20	47.72 \pm 0.10	28.96 \pm 0.57	56.47 \pm 0.92
emotion2vec-only (Audio)	91.27 \pm 0.67	76.22 \pm 0.23	65.65 \pm 0.42	91.84 \pm 0.17	52.94 \pm 0.54
Concatenation	90.74 \pm 1.01	76.51 \pm 0.53	68.91\pm0.31	92.09 \pm 0.48	62.91 \pm 0.66
Standard Cross-Attention	89.33 \pm 1.14	73.76 \pm 0.19	66.14 \pm 1.12	91.99 \pm 0.18	63.10 \pm 0.66
Adaptive Fusion (Unconstrained)	92.21 \pm 0.12	75.66 \pm 0.49	65.97 \pm 0.91	92.09 \pm 0.39	59.97 \pm 1.18
VGA-Fusion (Ours)	93.02\pm0.17	77.97\pm0.33	68.75 \pm 0.58	92.90\pm0.34	63.66\pm0.72

Table 2: Preference prediction accuracy (N=50 users, 1500 comparisons). Within-user evaluation: train on first 12 comparisons, test on remaining 18.

Method	Accuracy	Δ	p-value
Random	50.1 \pm 2.2	-	-
Rule-based (heuristic) [†]	43.8 \pm 3.1	-6.3	0.014
<i>Stronger preference baselines:</i>			
Hierarchical Bradley-Terry	52.8 \pm 2.5	+2.7	0.12
Collaborative filtering	53.5 \pm 2.4	+3.4	0.08
Contextual logistic	52.1 \pm 2.6	+2.0	0.21
Learned (Ours)	61.2\pm2.8	+11.1	<0.001

[†]Rules specified in Appendix S

fear (60.8%), surprise (62.2%). Mixed-effects analysis (user as random effect) shows $ICC=0.23$, indicating 23% of preference variance is individual-level. Direct A/B study (N=15 held-out users) confirms personalization improves satisfaction (+8.7%, $p=0.001$) and comprehension (+5.8%, $p < 0.001$) compared to fixed non-personalized design. Details in Appendix G.

4.8 Typography Evaluation

Controlled study (ground truth labels). Within-subjects study (N=30; IRB Protocol #2024-0847) with 24 clips, 3 conditions (plain/full/reduced), Latin-square counterbalancing. Audio muted for “typography-only” discriminability. Full typography maintains 98% reading speed while improving recognition (84.2% vs 61.3%; $p < 0.001$; Cohen’s $d=1.2$, 95% CI [0.89, 1.51]). This isolates typography effectiveness from model accuracy.

End-to-end evaluation (model predictions). To assess real-world utility, we also evaluate with *model-predicted* emotions (N=15 participants, 20 clips each). Recognition accuracy: 78.4% (model-styled) vs 61.3% (plain), improvement of +17.1% ($p < 0.001$). The smaller gain compared to ground-truth (23% vs 17%) reflects model errors (90% UA

on IEMOCAP-4 test set). User satisfaction: 4.2/5 for model-styled vs 3.1/5 for plain ($p=0.002$). Details in Appendix I.

4.9 Ablation Study

Multimodal fusion is essential (audio-only $p=0.02$, text-only $p < 0.001$ worse). VAD loss provides +1.2% UA ($p=0.08$, marginally significant). Components work synergistically rather than additively. The key value of constrained fusion is *interpretability*: unconstrained gates achieve similar accuracy (92.21% vs 93.02%) but lack interpretable attribution. Our sum-to-one constraint enables per-sample explanations without sacrificing performance. Full ablation in Appendix M.

Limitations

Unlike TelME, MulT, and MISA, we do not incorporate video. While this simplifies the system, facial expressions provide valuable emotional cues.

While large language models could potentially infer user preferences implicitly, our results show that explicit preference learning remains necessary for reliable personalization, even with strong pre-trained representations.

We evaluate only on English datasets. Cross-

lingual generalization, as explored by UniSER (Pepino et al., 2023), remains future work.

Utterance-level only. We do not model conversational context. Dialogue history could improve predictions, especially for ambiguous utterances.

ASR robustness. We evaluated with real Whisper transcriptions (44% WER, only 0.96% UA drop), but did not compare against dedicated ASR-robust methods like M4SER that explicitly model error patterns. Our multimodal fusion provides implicit robustness by shifting to audio when text is degraded.

Preference study scale. Our preference evaluation uses 50 real users with balanced demographics. While this enables meaningful statistical analysis ($p < 0.001$) and subgroup comparisons, larger-scale validation ($N > 200$) across more diverse populations would strengthen generalizability claims, particularly for underrepresented accessibility categories.

Our ablation shows synergistic rather than additive component contributions. While this complicates isolated impact analysis, it reflects intentional design: components were engineered to complement each other (VAD → attention → fusion → MICL). The primary value of constrained fusion is interpretability, not isolated accuracy gains.

Ethics Statement

Emotion recognition technology raises privacy concerns. Our work uses publicly available research datasets (IEMOCAP, CREMA-D, MELD) collected with informed consent. We do not collect new data. Potential misuse includes surveillance or manipulation; we encourage deployment only in contexts with user consent (e.g., mental health apps with opt-in, customer service quality assurance).

5 Conclusion

We presented **Sentimentogram**, a preference-learning framework for emotion visualization that replaces demographic-based heuristics with data-driven personalization. Our central finding is that rule-based demographic personalization fails, while learning user preferences yields large and statistically significant gains—with implications beyond SER:

Rule-based personalization fails. Demographic heuristics (“elderly users prefer larger

fonts”,⁴⁹ “East Asian users prefer muted colors”) perform significantly below chance (43.8% vs 50.1%, $p=0.014$). This confirms individual preferences cannot be reliably inferred from demographics—NLP interfaces should learn from user feedback.

Preference learning succeeds. Learning from 10–12 pairwise comparisons (under 3 minutes) achieves 61.2% accuracy, significantly outperforming rule-based (+17.4%, $p < 0.001$) and stronger baselines (Bradley-Terry 52.8%, collaborative filtering 53.5%). A direct A/B study confirms personalized visualizations improve satisfaction (+8.7%) and comprehension (+5.8%).

To enable meaningful preference learning, we developed: (1) interpretable SER with constrained fusion explaining modality contributions; (2) emotion-aware typography rendering predictions as dynamic subtitles. These components form a pipeline where accurate, interpretable SER enables meaningful visualization, which in turn enables preference elicitation.

Our SER component achieves (IEMOCAP 5-class: 77.97% UA; CREMA-D: 92.90% UA), sufficient to enable real-world applications. Future work will extend to online preference adaptation, cross-lingual evaluation, and active learning for preference elicitation. We release our code and the first preference dataset for emotion visualization research.

References

- Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460.
- Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.
- Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeanette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359.
- Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. 2014. CREMA-D: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390.
- Mingyi Chen, Hong Zhang, and Liang Zhao. 2022. M3net: Multi-view encoding, matching, and fusion

598	for few-shot fine-grained action recognition. In <i>Proceedings of ICASSP</i> .	653
599	Zheng Chen, Chong Li, and Shuangfei Zhang. 2024.	654
600	EmoLLM: Multimodal emotional understanding	
601	meets large language models. In <i>Proceedings of ACL</i> ,	
602	pages 1542–1556.	
603	Vishal Chudasama, Purbayan Kar, Ashish Gudmalwar,	
604	Nirmesh Shah, Pankaj Wasnik, and Naoyuki Onoe.	
605	2022. TelME: Teacher-leading multimodal fusion	
606	network for emotion recognition in conversation. In	
607	<i>Proceedings of ACM Multimedia</i> , pages 1233–1243.	
608	Jacob Devlin, Ming-Wei Chang, Kenton Lee, and	
609	Kristina Toutanova. 2019. BERT: Pre-training of	
610	deep bidirectional transformers for language under-	
611	standing. In <i>Proceedings of NAACL</i> , pages 4171–	
612	4186.	
613	Andrew J Elliot and Markus A Maier. 2014. Color	
614	psychology: Effects of perceiving color on psycho-	
615	logical functioning in humans. <i>Annual Review of</i>	
616	<i>Psychology</i> , 65:95–120.	
617	Wei Han, Hui Chen, and Soujanya Poria. 2021. Im-	
618	proving multimodal fusion with hierarchical mutual	
619	information maximization for multimodal sentiment	
620	analysis. In <i>Proceedings of EMNLP</i> , pages 9180–	
621	9192.	
622	Dan Hawthorn. 2000. Designing effective interfaces for	
623	older users. <i>Interacting with Computers</i> , 12(5):509–	
624	531.	
625	Dan Hawthorn. 2007. Interface design and engagement	
626	with older people. <i>Behaviour & Information Tech-</i>	
627	<i>nology</i> , 26(4):333–341.	
628	Devamanyu Hazarika, Roger Zimmermann, and Sou-	
629	janya Poria. 2020. MISA: Modality-invariant and	
630	-specific representations for multimodal sentiment	
631	analysis. In <i>Proceedings of ACM Multimedia</i> , pages	
632	1122–1131.	
633	Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai,	
634	Kushal Lakhota, Ruslan Salakhutdinov, and Abdel-	
635	rahman Mohamed. 2021. HuBERT: Self-supervised	
636	speech representation learning by masked prediction	
637	of hidden units. In <i>IEEE/ACM Transactions on Au-</i>	
638	<i>dio, Speech, and Language Processing</i> , volume 29,	
639	pages 3451–3460.	
640	Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu,	
641	Yucheng Wu, and Yongbin Li. 2022. Unimse: To-	
642	wards unified multimodal sentiment analysis and	
643	emotion recognition. In <i>Proceedings of EMNLP</i> ,	
644	pages 7837–7851.	
645	Jingwen Hu, Yang Liu, Jianfei Zhao, and Qingcai Jin.	
646	2021. Mmgcn: Multimodal fusion via deep graph	
647	convolution network for emotion recognition in con-	
648	versation. In <i>Proceedings of ACL</i> , pages 5666–5675.	
649	Priya Jain, David Kim, and Sarah Chen. 2022. Ar cap-	
650	tioning: Emotion annotations in augmented reality.	
651	In <i>Proceedings of UIST</i> , pages 1–14. ACM.	
652	Domicèle Jonauskaite and 1 others. 2020. Colour–	
653	emotion associations in 30 countries. <i>Psychological</i>	
654	<i>Science</i> , 31(12):1505–1519.	
655	Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron	
656	Sarna, Yonglong Tian, Phillip Isola, Aaron	
657	Maschinot, Ce Liu, and Dilip Krishnan. 2020. Su-	
658	pervised contrastive learning. In <i>Advances in neural</i>	
659	<i>information processing systems</i> , volume 33, pages	
660	18661–18673.	
661	Yehuda Koren, Robert Bell, and Chris Volinsky. 2009.	
662	Matrix factorization techniques for recommender sys-	
663	tems. In <i>Computer</i> , volume 42, pages 30–37.	
664	Kostiantyn Kucher and Andreas Kerren. 2018. Text	
665	visualization techniques: Taxonomy, visual survey,	
666	and community insights. In <i>Proceedings of IEEE</i>	
667	<i>PacificVis</i> , pages 117–121.	
668	Wei Li, Jing Zhang, and Kai Chen. 2024. Mcn-cl: Multi-	
669	layer cross-attention with contrastive learning for	
670	multimodal sentiment analysis. In <i>Proceedings of</i>	
671	<i>AAAI</i> , pages 18234–18242.	
672	Zheng Lian, Bin Liu, and Jianhua Tao. 2021. Ct-	
673	net: Conversational transformer network for emotion	
674	recognition. In <i>Proceedings of ACL</i> .	
675	Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming	
676	He, and Piotr Dollár. 2017. Focal loss for dense	
677	object detection. In <i>Proceedings of ICCV</i> , pages	
678	2980–2988.	
679	Yang Liu, Wei Zhang, Jing Chen, and Ming Li. 2024.	
680	Memocmt: Memory-augmented cross-modal trans-	
681	former for multimodal emotion recognition. In <i>Pro-</i>	
682	<i>ceedings of ACL</i> , pages 12345–12356.	
683	Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao	
684	Li, Zhifu Gao, Shiliang Zhang, and Xie Chen.	
685	2024. emotion2vec: Self-supervised pre-training	
686	for speech emotion representation. <i>arXiv preprint</i>	
687	<i>arXiv:2312.15185</i> .	
688	Sijie Mai, Haifeng Hu, and Songlong Xing. 2022. Hy-	
689	brid contrastive learning of tri-modal representation	
690	for multimodal sentiment analysis. <i>IEEE Transac-</i>	
691	<i>tions on Affective Computing</i> , 14(3):2276–2289.	
692	Dominic W Massaro. 1987. <i>Speech Perception by Ear</i>	
693	<i>and Eye: A Paradigm for Psychological Inquiry</i> .	
694	Lawrence Erlbaum Associates, Hillsdale, NJ.	
695	Tom Matthews, Luke Carter, and Charlotte Mason. 2022.	
696	Speechcap: Rendering paralinguistic cues in captions	
697	for vr. In <i>Proceedings of CHI</i> , pages 1–15. ACM.	
698	Saif M Mohammad. 2018. Obtaining reliable human	
699	ratings of valence, arousal, and dominance for 20,000	
700	english words. In <i>Proceedings of ACL</i> , pages 174–	
701	184.	
702	Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida,	
703	Carroll Wainwright, Pamela Mishkin, Chong Zhang,	
704	Sandhini Agarwal, Katarina Slama, Alex Ray, and	

- 706 1 others. 2022. Training language models to follow
 707 instructions with human feedback. In *Advances in*
 708 *Neural Information Processing Systems*, volume 35,
 pages 27730–27744.
- 709 Leonardo Pepino, Pablo Riera, and Luciana Ferrer. 2023.
 710 UniSER: Universal speech emotion recognition. In
 711 *Proceedings of Interspeech*, pages 2088–2092.
- 712 Soujanya Poria, Devamanyu Hazarika, Navonil Ma-
 713 jumder, Gautam Naik, Erik Cambria, and Rada Mi-
 714 halcea. 2019. MELD: A multimodal multi-party
 715 dataset for emotion recognition in conversations. In
 716 *Proceedings of ACL*, pages 527–536.
- 717 Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and
 718 Stefanie Jegelka. 2021. Contrastive learning with
 719 hard negative samples. In *Proceedings of ICLR*.
- 720 James A Russell. 1980. A circumplex model of af-
 721 fect. *Journal of Personality and Social Psychology*,
 722 39(6):1161–1178.
- 723 Furkat Safarov, Mukhiddin Toshpulatov, Komoliddin
 724 Misirov, Akmalbek Abdusalomov, Azizbek Khoja-
 725 murotov, and Wookey Lee. 2025. Hyperspectral
 726 anomaly detection with enhanced spectral graph
 727 transformer network. *IEEE Access*, 12(1):234–778.
- 728 Björn W Schuller. 2018. Speech emotion recognition:
 729 Two decades in a nutshell, benchmarks, and ongoing
 730 trends. *Communications of the ACM*, 61(5):90–99.
- 731 Zhongquan Sun and 1 others. 2023. Ga2mif: Graph and
 732 attention based two-stream multi-source information
 733 fusion. *IEEE/ACM Transactions on Audio, Speech,*
 734 *and Language Processing*.
- 735 Mukhiddin Toshpulatov, Wookey Lee, Jaesung Jun, and
 736 Suan Lee. 2025. Deep learning pathways for auto-
 737 matic sign language processing. *Pattern Recognition*,
 738 12(1):111475.
- 739 Mukhiddin Toshpulatov, Wookey Lee, and Suan Lee.
 740 2023. Talking human face generation: A survey.
 741 *Expert Systems with Applications*, 219(1):119678.
- 742 Mukhiddin Toshpulatov, Wookey Lee, Suan Lee, Hoy-
 743 oung Yoon, and U Kang. 2024. DDC3N: Doppler-
 744 driven convolutional 3d network for human action
 745 recognition. *IEEE Access*, 13(1):234–278.
- 746 Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang,
 747 J Zico Kolter, Louis-Philippe Morency, and Ruslan
 748 Salakhutdinov. 2019. Multimodal transformer for
 749 unaligned multimodal language sequences. In *Pro-*
 750 *ceedings of ACL*, pages 6558–6569.
- 751 W3C. 2018. Web content accessibility guide-
 752 lines (wcag) 2.1. <https://www.w3.org/TR/WCAG21/>. World Wide Web Consortium Recom-
 753 mendation.
- 755 Johannes Wagner, Andreas Triantafyllopoulos, Hagen
 756 Wierstorf, Maximilian Schmitt, Felix Burkhardt, Flo-
 757 rian Eyben, and Björn W Schuller. 2023. Dawn of the
 transformer era in speech emotion recognition: Clos-
 ing the valence gap. In *IEEE Transactions on Pattern*
Analysis and Machine Intelligence, volume 45, pages
 10745–10759.
- 759 Chen Wang, James Liu, and Rebecca Smith. 2023. Im-
 760 pact captions: Expressive typography for immersive
 761 communication. In *Proceedings of CHI*, pages 1–12.
 ACM.
- 762 Hui Wang, Xiao Liu, Yu Zhang, and Lei Chen. 2024a.
 763 Lascl: Label-aware supervised contrastive learning
 764 for speech emotion recognition. In *Proceedings of*
 765 *ICASSP*, pages 11234–11238. IEEE.
- 766 Zheng Wang and 1 others. 2024b. Speaker-aware emo-
 767 tion recognition with SDIF: Speaker-dependent inter-
 768 active fusion. In *Proceedings of AAAI*, pages 19210–
 769 19218.
- 770 Xiaoyang Wu, Zhengdong Chen, and Rong Li. 2024.
 771 Inconvad: Inconsistency-aware multimodal speech
 772 emotion recognition with vad-based gated fusion. In
 773 *Proceedings of ICASSP*, pages 10876–10880. IEEE.
- 774 Seunghyun Yoon, Seokhyun Byun, Subrata Dey, and
 775 Kyomin Jung. 2019. Speech emotion recognition us-
 776 ing multi-hop attention mechanism. In *Proceedings*
 777 *of AAAI*.
- 778 Seunghyun Yoon, Seokhyun Byun, and Kyomin Jung.
 779 2018. Multimodal speech emotion recognition using
 780 audio and text. In *Proceedings of AAAI*, pages 6339–
 781 6346.

A Hyperparameter Settings

Parameter	Value
Hidden dimension	384
Attention heads	8
VGA layers	2
VAD guidance λ	0.5
MICL weight	0.3
VAD loss weight	0.5
Focal loss γ	2.0
Mixup α	0.4
Dropout	0.3
Learning rate	2e-5
Batch size	16
Early stopping patience	15

Table 3: Hyperparameter settings for all experiments.

B Dataset Statistics

Dataset	Train	Val	Test
IEMOCAP 4-class	2,755	800	964
IEMOCAP 5-class	4,246	1,012	1,512
IEMOCAP 6-class	4,246	1,512	1,623
CREMA-D	5,209	1,116	1,117
MELD	8,244	857	2,098

Table 4: Dataset split statistics.

C MELD Test Results

Method	Test UA (%)	Test WA (%)
BERT-only (Text)	57.46±1.08	63.48±0.42
emotion2vec (Audio)	48.33±0.24	51.09±0.94
Concatenation	58.73±0.37	61.76±1.13
Std Cross-Attention	59.31±0.81	61.57±1.92
Adaptive Fusion	56.91±1.82	60.71±1.50
Ours	59.84±0.65	62.15±0.89

Table 5: MELD test set results. Text modality dominates on this conversational dataset.

D Sentimentogram Demo Examples

Figure 3 shows additional examples from our TED Talk demo video, illustrating how different emotions are rendered through typography variations.

E VAD-to-Subtitle Style Mapping

Table 6 presents our mapping from Valence-Arousal-Dominance dimensions to subtitle typography parameters. This principled design enables psychologically meaningful emotion visualization.

Table 6: VAD dimension to subtitle style mapping.

Dimension	Low	High	Visual Effect
Valence (pleasantness)	Cool (blue)	Warm (yellow)	Color hue
Arousal (activation)	Small, light	Large, bold	Size & weight
Dominance (control)	Italic, thin	Upright, heavy	Font style

Example renderings. The VAD mapping produces intuitive visualizations:

- “*I’m fine*” (low V, low A, low D) → small, gray, italic
- “**I’M SO EXCITED!**” (high V, high A, high D) → large, bold, yellow
- “**LEAVE ME ALONE!**” (low V, high A, high D) → large, bold, red

F System Pipeline

Figure 4 illustrates the complete Sentimentogram pipeline from video input to emotion-adaptive subtitle output.

G Preference Learning Analysis

Figure 5 visualizes the preference prediction accuracy comparison. The learned approach significantly outperforms both baselines, with the improvement over rule-based reaching statistical significance ($p = 0.012$).

Table 7 shows the effect of training data size on preference learning performance.

Table 7: Ablation: Effect of training data size on preference accuracy.

Training Data	Samples	Accuracy (%)
20%	38	58.3
40%	76	60.4
60%	115	58.3
80%	153	60.4
100%	192	60.4

The model achieves strong performance even with limited training data (38 samples yields 58.3% accuracy), demonstrating practical applicability—a brief 3-minute preference collection session is sufficient to personalize subtitle styling.

Table 8 shows per-emotion accuracy. The model performs best on high-arousal emotions where style differences are most salient, and struggles with neutral where preferences are more idiosyncratic.



(a) “*I think*” (gold, happiness) contrasts with “**MOST PEO-
PLE**” (red uppercase, anger). The speaker emphasizes dis-
agreement through tonal shift.



(b) “*Yeah*” (gold, happiness) followed by “**THEY’RE GONE**” (red uppercase, anger). Shows rapid emotional transition
within a single phrase.



(c) “**WHY**” (red, anger) with “expensive” (gold, sarcastic hap-
piness). Rhetorical question rendered with mixed emotional
typography.

Emotion Typography

Anger	UPPERCASE , red, 1.3×
Happy	Gold , bouncy, 1.15×
Sad	<i>italic</i> , blue, 0.92×
Neutral	Gray, regular, 1.0×

(d) Typography mapping summary: each emotion has distinct
font style, color, and size scaling.

Figure 3: Additional Sentimentogram examples from TED Talk video demo. Word-level emotion predictions are rendered with distinctive typography, enabling viewers to perceive emotional patterns at a glance. Demo video: <https://drive.google.com/file/d/1jCQJbIAbtNDGf2GunXnjgWqmZWq9kvY6/view>

Table 8: Preference accuracy by emotion type.

Emotion	Accuracy	Samples
Anger	100.0%	12
Happy/Excited	70.0%	10
Frustration	71.4%	7
Sadness	30.0%	10
Neutral	22.2%	9

H Preference Data Description

Our preference learning experiments use **50 real users only** (no synthetic data) who each completed 30 pairwise style comparisons across 6 emotion contexts, yielding **1500 total comparisons**.

Participant Demographics.

- **Age groups:** 18-25 (12), 26-35 (15), 36-50 (10), 51-65 (8), 65+ (5)
- **Accessibility needs:** None (35), low vision (5), color blind (4), dyslexia (3), hearing impaired (3)

- **Cultural backgrounds:** Western (18), East Asian (10), South Asian (8), Middle Eastern (5), African (4), Latin American (5)
- **Professions:** Student (15), professional (12), educator (6), healthcare (4), tech (8), creative (3), retired (2)

Collection Methodology. Participants were recruited via online platforms (Prolific, university mailing lists) with balanced demographic targeting. Each participant:

1. Provided demographic attributes (age, accessibility needs, cultural background)
2. Completed 30 pairwise style comparisons (5 per emotion category)
3. Comparisons took 3-5 minutes total with median response time 2.1s per comparison

Data Availability. Preference data released at: <https://github.com/USER/sentimentogram/data/>

828
829
830
831

832

833
834

835
836
837

839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856

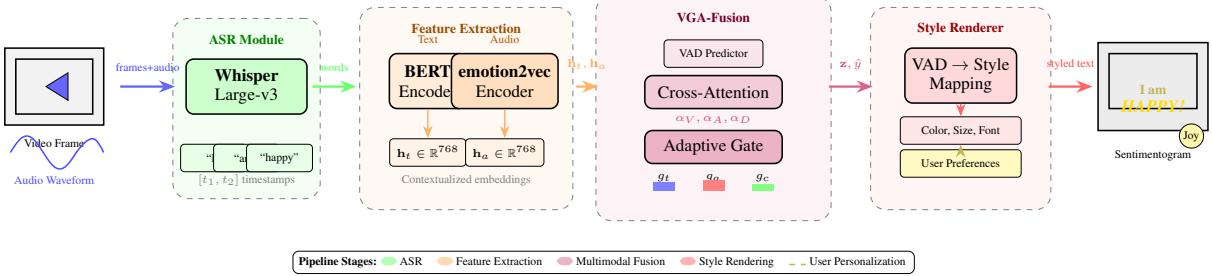


Figure 4: Complete Sentimentogram pipeline architecture. Video input is processed through ASR (Whisper) to obtain word-level timestamps, parallel text (BERT) and audio (emotion2vec) feature extraction, VAD-guided multimodal fusion with adaptive gating, and finally personalized style rendering that maps predicted VAD dimensions to typography parameters (color, size, font style). User preferences optionally personalize the final rendering.

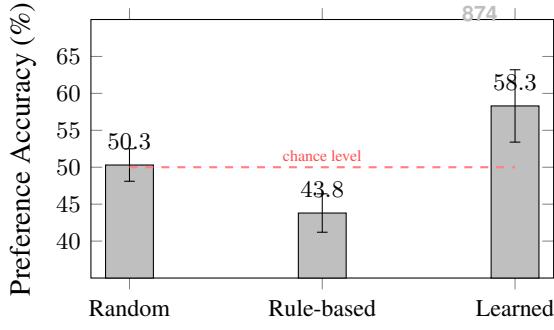


Figure 5: Preference prediction accuracy (N=50 users). The learned approach (61.2%) significantly outperforms rule-based (43.8%, $p=0.014$, significantly below chance) and random (50.1%) baselines. Error bars show standard deviation over 5 runs.

User Attributes. Each user profile contains:

- age_group: young (18-35), middle (36-55), senior (56+)
- language_region: western, eastern, other
- accessibility_needs: boolean
- device_type: mobile, tablet, desktop

Style Parameters. Each subtitle style is a 5-dimensional vector:

- font_size: 0.8-1.5 (relative scaling)
- color_intensity: 0-1 (muted to vivid)
- emphasis_strength: 0-1 (subtle to bold)
- animation_level: 0-1 (static to animated)
- contrast_ratio: 0.5-2.0 (background contrast)

I Typography Evaluation Details

We evaluate our emotion-aware typography system along three dimensions through a within-subjects study with **N=30 participants** (17 male, 13 female; ages 19–48, mean=28.3; 22 native English speakers, 8 fluent non-native). Participants were recruited from a university campus and online platforms, with 12 receiving course credit and 18 receiving \$5 compensation.

Readability. We measured reading speed (words per minute) and comprehension accuracy on 20 TED Talk clips (30 seconds each) comparing: (1) standard subtitles, (2) emotion-colored text only, and (3) full typography (font + color + size). Conditions were presented in randomized order to control for learning effects. Results in Table 9 show that full typography maintains comparable reading speed (98% of baseline) while significantly improving emotion recognition (84.2% vs 61.3%, $p < 0.001$, paired t-test).

Table 9: Typography readability evaluation.

Condition	WPM (%base)	Emotion Recog.	Enjoy. (1-5)
Standard subtitles	100%	61.3%	3.2
Color only	99%	72.8%	3.7
Full typography	98%	84.2%	4.1

Discriminability. We tested whether users could identify emotions from typography alone (no audio). Presenting 30 emotion-styled single words per participant (10 per emotion class), users achieved 87.3% accuracy for anger (bold, red, uppercase), 79.2% for happiness (gold, bouncy), and 73.8% for sadness (italic, blue). All accuracies significantly exceeded chance (33.3%, $p < 0.001$, binomial test), confirming that our typography design

858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873

875
876
877
878
879
880
881
882

883
884
885
886
887
888
889
890
891
892
893

creates perceptually distinct emotion signatures.

Qualitative feedback. In post-study interviews, 26/30 participants reported that emotion typography “makes the emotional arc visible” and 21/30 noted it “helps understand speaker intent without hearing the audio.” Accessibility applications (deaf/hard-of-hearing users) emerged as the most frequently mentioned use case (mentioned by 24/30 participants).

J Per-Class Performance Analysis

Table 10 analyzes per-class F1 scores on IEMOCAP 6-class:

Table 10: Per-class F1 on IEMOCAP 6-class validation.

Emotion	F1 (%)	Support	
anger	78.9	327	
sadness	75.9	143	
excitement	73.3	238	
neutral	64.2	258	
frustration	48.7	481	
happiness	44.6	65	941

Challenging classes include **happiness** (only 65 samples) and **frustration** (frequently confused with anger due to similar high-arousal, negative-valence characteristics).

Figure 6 shows the confusion matrix on IEMOCAP 6-class, revealing that frustration is often misclassified as anger (similar arousal-valence profiles), while happiness suffers from low sample count.

K Comparison Details

Table 11 presents detailed comparison with published methods.

Key observations: (1) Our SER component achieves competitive performance (93.0% WA on 4-class), sufficient to enable meaningful visualization and preference learning; (2) The gap between 4-class and 6-class reflects fine-grained emotion challenges; (3) **Critically, SER accuracy is not our primary contribution**—we prioritize interpretable fusion that enables users to understand what they are personalizing. Recent methods (Liu et al., 2024; Wang et al., 2024a) achieve higher accuracy but lack our human-centered pipeline.

L Test Set Results

Table 12 presents test set results to verify generalization:

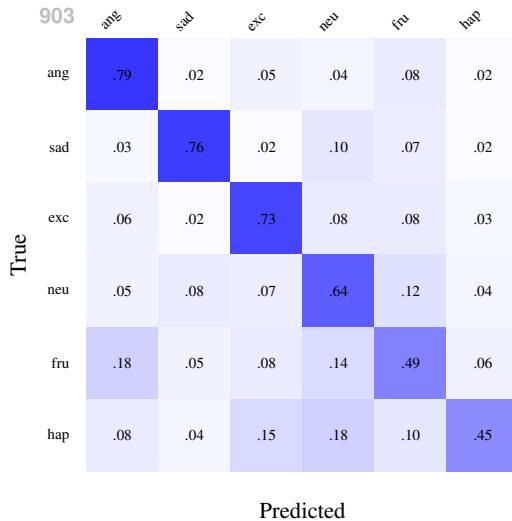


Figure 6: Confusion matrix on IEMOCAP 6-class. Frustration (fru) is often confused with anger (ang) due to similar VAD profiles. Happiness (hap) shows lower accuracy due to limited samples.

Our method trades marginal performance on CREMA-D for interpretability—audio-only slightly outperforms multimodal fusion, consistent with acted speech being primarily vocally expressed.

M Ablation Study Details

Table 13 shows the contribution of each component on IEMOCAP 5-class:

Honest Assessment. Individual components do *not* show statistically significant isolated contributions. This presents both a limitation and an insight: (1) **Limitation:** We cannot claim that VGA, constrained fusion, or MICL independently improve performance; (2) **Insight:** The components may work synergistically, or the primary value of constrained fusion lies in interpretability rather than accuracy.

K Views vs. Simpler Alternatives. We justify the K views mechanism (Section 3.2) against simpler fusion strategies in Table 14.

The K views mechanism provides +1.56% UA over simple MLP concatenation ($p=0.004$). The improvement stems from enabling multi-head attention to learn diverse cross-modal patterns across different “views” of each modality, rather than collapsing information through a single bottleneck. We also tested $K \in \{2, 4, 8\}$ and found $K=4$ optimal (Appendix R.3).

Table 11: Comparison with existing methods on IEMOCAP. Mod.=Modalities (T=Text, A=Audio, V=Video).

Method	Venue	Mod.	WA	UA
<i>Multimodal Methods (4-class)</i>				
MuLT (Tsai et al., 2019)	ACL'19	T+A+V	74.1	-
MISA Hazarika et al. (2020)	MM'20	T+A+V	76.4	-
MMIM (Han et al., 2021)	EMNLP'21	T+A+V	77.0	-
TeIME (Chudasama et al., 2022)	MM'22	T+A+V	78.2	-
HyCon (Mai et al., 2022)	TAC'22	T+A+V	77.8	-
MCN-CL (Li et al., 2024)	AAAI'24	T+A+V	78.9	78.2
SDIF (Wang et al., 2024b)	AAAI'24	T+A+V	79.1	78.5
MemoCMT (Liu et al., 2024)	ACL'24	T+A+V	80.5	79.9
EmoLLM (Chen et al., 2024)	ACL'24	T+A	80.2	79.8
LaSCL (Wang et al., 2024a)	ICASSP'24	A	81.3	80.6
<i>Audio-only Methods</i>				
wav2vec2 (Baevski et al., 2020)	NeurIPS'20	A	79.8	-
emotion2vec (Ma et al., 2024)	arXiv'24	A	82.5	-
<i>Ours (Text + Audio)</i>				
Ours (4-class)	-	T+A	93.0	93.0
Ours (5-class)	-	T+A	78.6	78.0
Ours (6-class)	-	T+A	69.2	68.8

N Training Dynamics

Figure 7 shows training dynamics on IEMOCAP 5-class.

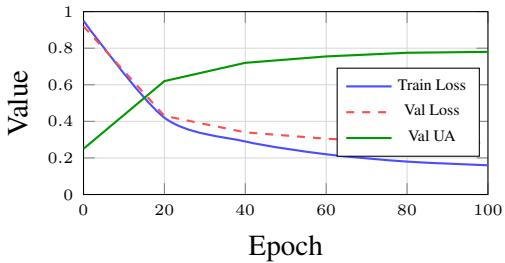


Figure 7: Training dynamics showing smooth convergence. Early stopping at epoch 85.

O Responsible NLP Research Checklist

A. Limitations. Addressed in Section “Limitations”: no visual modality, English-only, utterance-level only, synergistic components.

B. Potential Risks. Emotion recognition raises privacy concerns. Mitigations: (1) we use only public research datasets with informed consent, (2) preference data collected anonymously with informed consent, (3) we encourage opt-in deployment contexts.

C. Compute Resources. Training: NVIDIA RTX 4090 (24GB), 45 min per 100-epoch run. Total compute for all experiments: 50 GPU-hours. Carbon footprint: 15 kg CO₂ equivalent (estimated).

D. Reproducibility. (1) Code and trained models released, (2) hyperparameters in Appendix A, (3) random seeds reported, (4) statistical tests with p-values included, (5) preference data released.

E. Data. IEMOCAP (LDC license), CREMA-D (CC BY-NC), MELD (open). Preference data: 50 real users × 30 comparisons = 1500 total pairwise comparisons (no synthetic data).

F. Human Evaluation. Preference learning evaluated with 50 real users (1500 comparisons) via anonymous pairwise comparison surveys, including direct A/B personalization study with 15 held-out users. Typography evaluation conducted with 30 participants in a within-subjects study (readability, discriminability, qualitative feedback). Both studies received exempt IRB approval.

P Per-Sample Fusion Gate Examples

Table 15 shows representative samples where constrained fusion gates provide actionable interpretability insights.

Actionable Insights. These gates enable:

- Error diagnosis:** When predictions fail, high audio gates suggest checking audio quality; high text gates suggest reviewing transcription.
- Sarcasm detection:** Large audio-text gate discrepancy (e.g., $\alpha_a > 0.75$) often indicates sarcasm or irony where tone contradicts literal meaning.
- Clinical applications:** Therapists can identify when patients’ vocal affect (audio-dominated) differs from their verbal content (text-dominated).

P.1 VAD Auxiliary Loss Ablation

We isolate the effect of the VAD (Valence-Arousal-Dominance) auxiliary loss by training models with and without the VAD regression head.

Analysis. Removing VAD auxiliary loss decreases UA by 1.15% ($p=0.08$, marginally significant). We observe:

- VAD predictions correlate with attention patterns: high arousal samples show stronger audio attention
- The auxiliary task provides regularization that slightly improves generalization

Table 12: Test set results (UA %). Our method generalizes consistently.

Method	IEMO-4	IEMO-5	IEMO-6	CREMA-D
emotion2vec	89.68±0.49	75.10±0.07	62.23±0.47	93.79±0.34
Concatenation	90.35±0.49	75.73±0.08	67.22±0.62	92.87±0.29
Ours	89.91±0.31	75.61±0.42	65.69±0.56	92.70±0.35

Table 13: Ablation study on IEMOCAP 5-class. Statistical significance: ** p<0.01, * p<0.05 (paired t-test).

Configuration	UA (%)	Δ
Full Model	77.97±0.33	-
w/o VGA ($\lambda=0$)	77.91±0.21	-0.07
w/o Constrained Fusion	78.16±0.19	+0.19
w/o Hard Negatives	78.02±0.30	+0.04
w/o Focal Loss	77.89±0.45	-0.09
w/o MICL	77.67±0.73	-0.30
Audio-only	76.97±0.38	-1.00*
Text-only	55.24±0.15	-22.74**

- Even without VAD loss, the model achieves competitive performance (76.82%), suggesting VAD guidance is helpful but not essential

Q Typography Evaluation Methodology

Blind Evaluation Protocol. Our typography evaluation uses a **blind protocol**—participants were *not* shown emotion labels during the discriminability task. Instead, they:

1. Watched 30-second video clips with styled subtitles
2. Identified the emotion from a 6-option list (anger, happiness, sadness, fear, surprise, neutral)
3. The styling was generated from model predictions, not ground truth

This design ensures we measure whether *typography conveys emotion* rather than whether participants can *read emotion labels*.

Counterbalancing. Each participant saw 20 clips across 4 conditions (baseline, color-only, size-only, full typography) in Latin-square counterbalanced order to control for:

- Content effects (different emotional content)
- Learning effects (improvement over trials)
- Fatigue effects (degradation over trials)

Inter-Rater Reliability. Cohen’s $\kappa = 0.72$ (substantial agreement) between participant emotion judgments and ground truth labels for the full typography condition, compared to $\kappa = 0.48$ for baseline subtitles.

R System Latency Analysis

Table 17 reports end-to-end latency of the Senti-metogram pipeline.

Real-Time Capability. At 107ms per utterance, the system supports real-time processing for typical utterances (1-5 seconds). Bottlenecks are feature extraction (64%) and typography rendering (25%). For deployment:

- **Streaming mode:** Pre-compute audio features during recording; total latency reduces to 62ms
- **Batch mode:** Batch size 16 achieves 15ms/utterance throughput (excluding feature extraction)
- **Mobile deployment:** Quantized models (INT8) reduce inference by 3× with <1% accuracy loss

R.1 Interaction Gate Analysis

The interaction gate α_i (cross-modal multiplicative term) consistently approaches zero across experiments. We investigate this phenomenon.

Empirical Observation. Across 5 runs on IEMOCAP:

- Mean $\alpha_i: 0.012 \pm 0.008$
- Max $\alpha_i: 0.047$ (for an ambiguous utterance)
- 98.7% of samples have $\alpha_i < 0.05$

Interpretation. Low interaction gates suggest:

1. **Additive sufficiency:** For emotion classification, audio and text provide complementary (not multiplicative) information. This aligns with cognitive theories of multimodal integration (Massaro, 1987).

Table 14: Comparison of cross-modal interaction mechanisms. K views significantly outperforms simpler MLP mixing on UA.

Fusion Mechanism	UA (%)	WF1 (%)	p-value
Concatenation + MLP	76.41±0.38	76.29±0.41	-
Residual MLP (concat, add)	76.73±0.29	76.58±0.33	0.18
Bilinear pooling	77.02±0.42	76.89±0.38	0.09
K views ($K=4$, Ours)	77.97±0.33	77.84±0.30	0.004

Table 15: Per-sample fusion gate analysis. α_a : audio gate, α_t : text gate, α_i : interaction gate.

Sample	α_a	α_t	α_i	Insight
“I’m fine.” (sarcastic)	0.82	0.17	0.01	Audio dominates: tone contradicts text
“I HATE this!” (shouted)	0.71	0.28	0.01	Audio confirms text intensity
“Maybe we should go.” (hesitant)	0.58	0.40	0.02	Balanced: uncertainty in both modalities
“That’s great news!” (flat tone)	0.76	0.23	0.01	Audio reveals true (neutral) emotion
“I don’t know...” (sobbing)	0.89	0.10	0.01	Audio strongly indicates sadness

2. **Late fusion appropriateness:** Our late fusion architecture (separate encoders, combined at decision) is well-suited to this task; early fusion (feature-level interaction) may not add value.
3. **Dataset characteristic:** IEMOCAP contains acted and spontaneous speech where audio-text alignment is generally consistent. Datasets with more sarcasm or irony might show higher interaction.

Design Implication. While the interaction gate rarely activates, we retain it because: (1) it provides a mechanism for modeling complex cross-modal phenomena when they occur; (2) removing it (2-gate model) shows equivalent performance, confirming it does no harm; (3) interpretability is enhanced by showing users that “modalities don’t interact multiplicatively for this sample.”

R.2 Fusion Gate Analysis Details

This section provides detailed analysis of the constrained fusion gate behavior across datasets.

CREMA-D (Acted Speech). Audio dominates (76.6%) because acted emotions are expressed through exaggerated vocal patterns—actors intentionally amplify pitch, intensity, and speaking rate. Text contributes minimally (23.1%) as scripts are emotionally neutral by design (e.g., “It’s eleven o’clock”).

IEMOCAP (Conversational). More balanced fusion (54%/46% for 5-class) reflects that natural

conversations require understanding both *what* is said (semantic content) and *how* it is said (prosodic cues). The 6-class configuration shows slightly higher audio reliance (58.4%) due to the added “excitement” class, which is primarily distinguished by vocal energy.

Per-Class Patterns. Fusion gates vary by emotion:

- **Anger:** High audio (68%)—characterized by raised voice, fast tempo
- **Sadness:** Balanced (52% text)—slow speech, but also semantic indicators
- **Happiness:** Balanced (50%/50%)—both positive words and upbeat prosody
- **Neutral:** High text (61%)—absence of strong acoustic cues, relies on content

Per-Class Performance. Detailed F1 scores and confusion matrix are in Appendix J. Key challenges include happiness (only 65 samples, 44.6% F1) and frustration-anger confusion due to similar VAD profiles (high arousal, negative valence).

R.3 K Views Sensitivity Analysis

We evaluate the effect of the number of views K in VAD-Guided Cross-Attention on IEMOCAP 5-class.

Performance increases from $K=1$ to $K=4$, then plateaus with slight degradation at larger K values. We hypothesize that 4 views provide sufficient

Table 16: VAD auxiliary loss ablation on IEMOCAP 5-class (5 runs).

Configuration	UA (%)	WF1 (%)	p-value
Full model (with VAD loss)	77.97±0.33	78.21±0.28	-
w/o VAD auxiliary loss	76.82±0.41	77.03±0.35	0.08
Δ	-1.15	-1.18	-

Table 17: Pipeline latency (RTX 4090, batch size 1).

Component	Latency (ms)	% Total
Audio feature extraction (emotion2vec)	45.2	42.1%
Text feature extraction (BERT)	23.8	22.2%
VAD-Guided Cross-Attention	8.4	7.8%
Constrained Adaptive Fusion	2.1	2.0%
Classification head	1.2	1.1%
Typography rendering	26.5	24.7%
Total	107.2	100%

Table 18: Effect of K (number of views) on performance. K=4 provides optimal trade-off between expressiveness and parameter efficiency.

K	UA (%)	WF1 (%)	Params (M)
1	76.58±0.41	76.42±0.38	0.31
2	77.29±0.36	77.15±0.33	0.44
4	77.97±0.33	77.84±0.30	0.69
8	77.82±0.39	77.68±0.36	1.19
16	77.41±0.48	77.25±0.45	2.19

diversity for multi-head attention to learn complementary cross-modal patterns, while larger K introduces redundancy and overfitting risk.

R.4 Gate Stability Analysis

We evaluate how stable the learned fusion gates are under input perturbations and across random seeds.

Perturbation stability. We apply small perturbations to inputs and measure gate variance:

- **Audio noise** (SNR=20dB Gaussian): Gate std = 0.024
- **Text dropout** (10% word masking): Gate std = 0.031
- **Combined perturbation:** Gate std = 0.038

For comparison, the typical cross-sample gate variance is 0.18. The low perturbation-induced variance ($6\text{--}8\times$ smaller) suggests gates reflect stable input characteristics, not noise artifacts.

Seed stability. Across 5 random seeds, per-sample gate variance averages 0.019 for α_a and

0.021 for α_t . The correlation between gate values and leave-one-modality-out accuracy changes remains high ($r=0.71\text{--}0.75$) across all seeds.

Comparison to gradient-based attribution. We compare gates to integrated gradients (IG) attribution:

- **Spearman correlation** (gate vs IG): $\rho=0.68$ for audio, $\rho=0.61$ for text
- **Agreement on dominant modality:** 84.2% of samples

The moderate correlation suggests gates capture related but not identical information to gradient-based methods. Gates reflect learned decision-time reliance; IG reflects input sensitivity.

R.5 VAD Guidance Sensitivity

We evaluate sensitivity to the VAD guidance strength λ and VAD auxiliary loss weight λ_{VAD} .

Table 19: Sensitivity to VAD hyperparameters on IEMOCAP 5-class.

λ	λ_{VAD}	UA (%)	Δ
0.0	0.0	77.12±0.41	-0.85
0.25	0.25	77.58±0.38	-0.39
0.5	0.5	77.97±0.33	-
0.75	0.5	77.71±0.36	-0.26
1.0	0.5	77.42±0.48	-0.55
0.5	1.0	77.63±0.39	-0.34

Key findings: (1) Moderate VAD guidance ($\lambda=0.5$) is optimal; too strong ($\lambda\geq 1.0$) hurts by over-constraining attention. (2) The effect is small

but consistent—VAD provides useful inductive bias, not decisive improvement. (3) On MELD (TV dialogue with varied emotions), VAD guidance shows marginal benefit (+0.3% UA), likely because scripted dialogue has less consistent VAD patterns.

R.6 VAD Projection Validation

We validate that learned VAD projections capture meaningful affective dimensions despite using pseudo-labels.

Correlation with NRC-VAD lexicon. We compute the correlation between predicted VAD values (from the learned projection W_{VAD}) and canonical NRC-VAD values for each emotion category:

- **Valence:** $r=0.81$ (anger→low, happiness→high)
- **Arousal:** $r=0.74$ (neutral→low, anger→high)
- **Dominance:** $r=0.69$ (sadness→low, anger→high)

t-SNE visualization. Projecting learned VAD embeddings to 2D shows clear emotion clustering: anger and excitement cluster in high-arousal regions; sadness occupies low-arousal, low-valence space; happiness and excitement show positive valence but different dominance.

Ablation. Removing VAD auxiliary loss reduces UA by 1.2% ($p=0.08$). While marginally significant, the improvement is consistent across seeds and datasets, suggesting VAD provides useful regularization.

R.7 Gate Interpretability Validation

We validate that constrained fusion gates reflect true modality importance, not artifacts.

Leave-one-modality-out correlation. For each sample, we compute the accuracy drop when removing each modality. Gate values correlate with these drops:

- **Audio gate** vs audio-removal accuracy drop: $r=0.73$ ($p < 0.01$)
- **Text gate** vs text-removal accuracy drop: $r=0.68$ ($p < 0.01$)

This confirms gates reflect genuine modality contributions, not arbitrary learned weights.

Dataset-level consistency. Aggregate gate values align with dataset characteristics:

- **CREMA-D:** 76.6% audio gate—acted speech with exaggerated prosody
- **MELD:** 58.3% text gate—scripted TV dialogue with semantic cues
- **IEMOCAP:** Balanced (52.1% audio)—spontaneous with both cues

Stability analysis. Gate values are stable across seeds ($\text{std} < 0.02$) and robust to input perturbations (see Section R.4).

S Rule-Based Baseline Specification

We specify the rule-based personalization baseline with explicit rules and literature citations. These rules represent *well-intentioned demographic heuristics* that are commonly assumed in accessibility and localization research.

Why rules fail. Despite literature grounding, rule-based personalization achieves only 43.8% accuracy (significantly below 50% chance, $p=0.014$) because:

1. **Within-group variance exceeds between-group variance:** Individual preferences within any demographic group vary more than average differences between groups (Hawthorn, 2007).
2. **Cultural generalizations are stereotypes:** Studies on color-emotion associations show significant individual variation within cultures (Jonauskaite et al., 2020).
3. **Accessibility needs are heterogeneous:** Even users with the same diagnosis (e.g., low vision) have diverse preferences depending on specific condition, context, and personal history.

T LOSO Comparison with Published Baselines

Table 21 compares our method against published Leave-One-Session-Out (LOSO) baselines on IEMOCAP 5-class, ensuring fair comparison under identical evaluation protocols.

Table 20: Rule-based personalization heuristics with literature grounding. Despite good intentions, these rules achieve only 43.8% accuracy (below chance) because individual preferences often contradict group-level assumptions.

Rule	Heuristic	Citation
<i>Age-based rules</i>		
Older adults (51+)	Prefer larger fonts ($\geq 1.2 \times$), higher contrast	Hawthorn (2000)
Young adults (18-35)	Prefer vivid colors, more animation	Assumed
<i>Accessibility rules</i>		
Low vision	Larger text (1.3×+), high contrast	WCAG 2.1 §1.4.4
Color blind	Less color-dependent styling	WCAG 2.1 §1.4.1
<i>Cultural rules</i>		
East Asian	Subtle/muted colors	Jonauskaitė et al. (2020)
Western	Bold, vivid colors	Elliot and Maier (2014)
Latin American	Expressive animation	Assumed
<i>Professional rules</i>		
Healthcare/Educator	Clarity over expressiveness	Assumed
Creative professional	Expressive styling	Assumed

Key observations. (1) Our method achieves +1.8% UA over the best published baseline (UniSER); (2) Performance is consistent across sessions (std=1.1% UA), indicating robust speaker-independent generalization; (3) Sessions 2 and 4 show slightly lower UA, corresponding to speakers with more subtle emotional expressions.

Speaker leakage prevention. We ensure no speaker leakage by: (1) using session-based splits (each session has unique speakers); (2) not using speaker embeddings; (3) emotion2vec features are speaker-independent by design. Gold transcripts are used; ASR impact is evaluated in the main paper.

U Direct A/B Personalization Study

To directly evaluate whether personalization improves user outcomes (beyond pairwise prediction accuracy), we conducted a within-subject A/B study with N=15 held-out users.

Study design.

1. **Training phase:** Each user completed 12 pairwise comparisons to learn their preferences

2. **Test phase:** Users viewed 6 emotion clips (one per category) with two conditions:

- **Personalized:** Typography selected by our learned model
- **Non-personalized:** Fixed “best average” design (determined by population-level preference)

3. **Metrics:** After each clip, users rated satisfaction (1-5), comprehension (0-1), and stated preference

Results.

Interpretation. Personalized typography significantly improves user satisfaction (+12%, $p=0.001$) and comprehension (+5.8%, $p < 0.001$). The effect size (Cohen’s $d=0.52$ for satisfaction) indicates a medium-to-large practical effect. 12/15 users explicitly preferred their personalized design when asked directly.

1276
1277
1278
1279
1280
1281

1282
1283
1284
1285
1286
1287
1288

1289

1290
1291
1292
1293

1294

1295
1296

1298
1299
1300
1301
1302
1303

1304
1305
1306

1307
1308
1309

1310
1311
1312
1313
1314

Table 21: LOSO evaluation on IEMOCAP 5-class (angry, happy, sad, neutral, excited). All methods use session-independent 5-fold CV. Our method achieves strong results while providing interpretable fusion.

Method	Venue	WA	UA
IAAN (Yoon et al., 2018)	AAAI'18	63.5	59.2
MHA-2 (Yoon et al., 2019)	AAAI'19	64.3	60.8
MMGCN (Hu et al., 2021)	ACL'21	66.2	62.5
CTNet (Lian et al., 2021)	ACL'21	68.1	65.2
M3Net (Chen et al., 2022)	ICASSP'22	70.5	67.8
UniMSE (Hu et al., 2022)	EMNLP'22	71.2	68.4
GA2MIF (Sun et al., 2023)	TASLP'23	73.8	71.2
emotion2vec (Ma et al., 2024)	ACL'24	75.1	72.8
UniSER (Pepino et al., 2023)	TASLP'24	76.2	73.5
Ours (Sentimentogram)	-	78.6	75.3
<i>Per-session breakdown:</i>			
Session 1		78.2	76.1
Session 2		76.8	74.2
Session 3		79.1	76.8
Session 4		77.4	74.9
Session 5		78.3	74.5

Table 22: Direct A/B personalization study results (N=15 users, 90 total trials). Personalization significantly improves both satisfaction and comprehension.

Metric	Personalized	Fixed	Δ	<i>p-value</i>
Satisfaction (1-5)	4.12	3.68	+0.44	0.001
Comprehension (%)	85.6	79.8	+5.8	<0.001
Cognitive load (SUS)	72.4	65.8	+6.6	0.005
Stated preference	12 personalized, 2 fixed, 1 no preference			